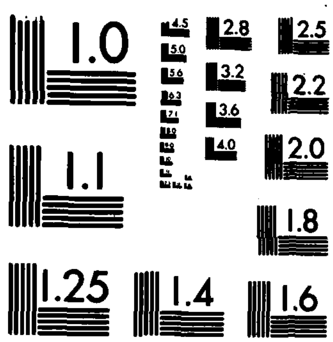
AD-A160 960   DESIGN EFFECTS OF TWO-STAGE SAMPLING(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER C J SKINNER JUL 85 MRC-TSR-2841 DAAG29-80-C-0041   1/1

UNCLASSIFIED   F/G 12/1   NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(2)

AD-A160 960

MRC Technical Summary Report #2841

DESIGN EFFECTS
OF TWO-STAGE SAMPLING

C. J. Skinner

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

July 1985

(Received June 27, 1985)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

85

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

DESIGN EFFECTS OF TWO-STAGE SAMPLING

C. J. Skinner

Technical Summary Report #2841

July 1985

ABSTRACT

The effect of a two-stage sampling design on statistical inference is discussed. A definition of a design effect is given. The structure of design effects for a class of statistics is investigated. Results have both a design-based and a model-based interpretation. The relation between design effects for multivariate statistics and design effects for univariate statistics is considered.

AMS (MOS) Subject Classifications: 62D05, 62H10

Key Words: Design Effect; Model Misspecification; Two-stage Sampling; Finite Population; Sample Survey.

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. DMS-8210950, Mod. 1.

SIGNIFICANCE AND EXPLANATION

In sample surveys, the design effect of a statistic is usually defined as the ratio of its true variance under the given sample design to its variance had the sample been obtained by simple random sampling.

Empirical work suggests certain patterns for design effects of different types of statistics under different designs but theoretical work explaining these patterns is limited. This paper obtains general theoretical results on the structure of design effects for a broad class of ~~statistics~~ (statistical inference) under a two-stage sampling design. In particular, it discusses the relation between design effects of multivariate and of univariate statistics.

This relation is of practical interest because it is of relevance to the imputation of standard errors for multivariate statistics such as correlation coefficients or regression coefficients using design effects of univariate statistics. The latter quantities are often routinely derived on completion of the survey. The former may be difficult to compute by standard procedures, either because of the absence of the necessary design information or because of software or degrees of freedom limitations.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

DESIGN EFFECTS OF TWO-STAGE SAMPLING

C. J. Skinner

1. INTRODUCTION

The application of statistical methods such as regression analysis and multivariate analysis to sample survey data is now widespread. Such methods typically assume that the rows of an $n \times q$ data matrix, x_n, are realizations of independent and identically distributed (IID) random vectors. A general question may therefore be raised as to the validity of inference procedures which make this assumption when the data is derived using a complex sample design. In particular, this paper is concerned with the effect of two-stage sampling on the estimation of functions of population moments, such as correlation coefficients.

The term 'design effect' was originally introduced (Kish, 1965) as a measure of efficiency for comparing sample designs. More recently (e.g. Rao and Scott, 1981) it has also been used as a measure of the impact of a sample design on an inference procedure. We shall be concerned only with this latter concept.

We presume a basic acquaintance with the distinction between the design-based and the model-based approaches to survey-sampling inference (e.g. Sarndal, 1978). From the design-based viewpoint the interpretation of 'the effect of two-stage sampling' is clear. The IID assumption corresponds to the randomization distribution induced in x_n by simple random sampling with replacement from a finite population (or without replacement from an infinite population). Two-stage sampling induces a different distribution in x_n and consequently perturbs the distribution of estimators from that predicted by IID theory.

From the model-based viewpoint the effect of the sampling design on inference is much less clear. The model-based approach begins by specifying a model distribution for the matrix of values, x, of the population units. Inference then proceeds in one of two

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. DMS-8210950, Mod. 1.

ways:

(A) Inference is based only on the model distribution conditional on the units actually obtained in the sample and irrespective of any other sample that might have been selected.

(B) Inference incorporates both the model distribution and the randomization distribution induced by the sample design.

The role of the sample design in model-based inference is by no means a subject of universal agreement (see, for example, the discussion of Royall and Cumberland, 1981). Sugden and Smith (1984) specify various conditions for choosing between (A) and (B). For example, an instance when it may be inappropriate to ignore the sample design occurs when sampling on the dependent variable in regression analysis (c.f. Nathan and Holt, 1980). In such cases the design has a direct effect on inference.

In Section 2 we adopt procedure (A). In this case the effect of the sample design is more indirect. For example, two-stage sampling presumes that the population is divided into clusters. Units within clusters usually tend to be more alike than units in different clusters. This implies that the IID assumption for x_n corresponds to an inappropriate model assumption. The 'effect of the design' is therefore really the effect of misspecifying the model (c.f. Scott and Holt, 1982, p. 850). For if the true model for x_n is in fact IID then two-stage sampling would have no effect under procedure (A). Conversely, if the true model is not IID and we happen to choose the same sample of units by (i) simple random sampling and (ii) two-stage sampling then the effect on inference is identical for (i) and (ii).

Our approach will be to define a distribution for x_n which has both a design-based and model-based interpretation and then to obtain results which may be interpreted as respectively design effects or misspecification effects. Because of the mathematical isomorphism between the results under the two approaches it will be convenient to use the single term design effect. We maintain, however, that this effect has distinct interpretations under the two approaches.

There is a further problem from the model-based viewpoint with the effect of design on statistical methods such as regression analysis. Suppose we take a two-stage sample and decide that the appropriate model allows for different regression relationships in different clusters. It may be argued (e.g. Pfefferman and Nathan, 1981) that the target parameters of interest are then the individual cluster regression coefficients rather than any overall population regression coefficient. We shall ignore this consideration here and assume that the target is a well defined population parameter. We view the design as an arbitrary selection process with no characteristic of substantive interest upon which we wish to 'condition' (c.f. Kish and Frankel, 1974).

We now introduce our basic definition of 'design effect'. We take x_n as a member of the infinite sequence $\{x_n; n=1,2,\ldots\}$. Let $\pi_o = \pi_{0\ n}$ be the 'baseline' distribution of x_n under the IID assumption. Let $\pi_1 = \pi_{1,n}$ be the true distribution of x_n. From the design-based viewpoint π_1 is the randomization distribution induced by the complex sampling design. From the model-based viewpoint, assuming procedure (A) above, π_1 is the true model distribution of x 'marginalized' to x_n.

Definition 1.1: Suppose $t_n = t_n(x_n)$ is a scalar statistic obeying the following central limit laws as $n \to \infty$.

$$n^{1/2}(t_n-\theta_0) \xrightarrow{L} N(0,\sigma_0^2) \quad \text{under } \pi_0$$

$$n^{1/2}(t_n-\theta_1) \xrightarrow{L} N(0,\sigma_1^2) \quad \text{under } \pi_1$$

Suppose also that $v_{0,n} = v_{0,n}(x_n)$ is consistent for σ_0^2 under π_0 and converges in probability to $\text{plim}_{\pi_1}(v_{0,n})$ under π_1. Then the design effect of t_n is defined as

$$\text{deff}(t_n;\pi_1,v_{0,n}) = \sigma_1^2/\text{plim}_{\pi_1}(v_{0,n}) \quad . \tag{1.1}$$

Remarks

1. The traditional definition of a design effect (e.g. Kish, 1965, p. 265) as a measure of design efficiency is σ_1^2/σ_0^2 in the above notation. Definition 1.1 is more natural as a measure of the impact of the design on estimation. It measures the effect of acting as if π_0 is true when in fact π_1 is true. Note that $\text{deff}^{1/2}$

provides a multiplicative adjustment for the standard-error estimate, $(n^{-1}v_{0,n})^{1/2}$.

2. The design effect will usually be of secondary importance if t_n is inconsistent under π_1, that is if θ_1 is not the target parameter. If π_0 is assumed to be the true distribution then t_n is usually chosen such that θ_0 is the target parameter. Then $\theta_1 - \theta_0$ is the asymptotic bias.

3. Definition 1.1 does not depend on π_0, except as a heuristic device for deriving $v_{0,n}$. This makes this definition easier to use from the model-based approach than the definition σ_1^2/σ_0^2.

4. The design effect is unity when $\pi_0 = \pi_1$.

5. The asymptotic nature of the definition simplifies results but is not essential.

In this article we shall be interested in how design effects depend upon the survey design and upon the population. We adopt a theoretical approach as opposed to the empirical approach of, for example, Kish and Frankel (1974). The latter approach may be more realistic but lacks generalizability because of the enormous range of possible statistics and population structures. The theoretical approach must make strong assumptions to obtain useful results but the extent of possible generalization should be more apparent. Of course the two approaches should complement each other.

We shall be particularly interested in the relation between design effect of multivariate statistics and design effects of univariate statistics. Such relations are of practical interest for at least two reasons. Firstly, the survey data collection agency may publish design effects for univariate statistics but, for confidentiality reasons, may not make sufficient survey design information available on public use tapes for the data analyst to estimate standard errors in the usual way. Given suitable theoretical relations, the analyst could instead impute standard error for multivariate statistics using the published univariate design effects. Secondly, even if the analyst has available full design information it may still be desirable to impute standard errors because of computer software availability or degrees of freedom limitations.

In Section 2 we outline the basic formal framework and define π_0 and π_1. In Section 3 we apply Definition 1.1 to a general class of estimators under the given π_1 and derive results on the form of design effects for the case of equal cluster sizes. An example is given in Section 4. The case of unequal cluster sizes is considered briefly in Section 5 and the implications of the results are discussed in Section 6.

2. FRAMEWORK AND ASSUMPTIONS

Consider a finite population, U, partitioned into K clusters. Let the j^{th} unit in the i^{th} cluster be labelled (i,j) for $i = 1,\ldots,K$, $j = 1,\ldots,M_i$ where M_i is the size of the i^{th} cluster. A sample is a subset, S, of $U = \{(i,j) \, ; \, i = 1,\ldots,K, j = 1,\ldots,M_i\}$. We suppose that the sample is selected in such a way that each subset S of U has a known probability, $p(S)$, of selection. Conventionally the sample is chosen in two stages: first, a sample of clusters is selected and then subsamples are selected within each of the selected clusters. Without loss of generality we write the actual sample obtained as $s = \{(i,j) \, ; \, i = 1,\ldots,k, j = 1,\ldots,m_i\}$. The sizes of the sample and population are respectively:

$$n = \sum_1^k m_i \quad , \quad N = \sum_1^K M_i \quad .$$

We suppose that a $q \times 1$ vector x_{ij} is associated with unit (i,j) in U and let

$$x = (x_{11},\ldots,x_{KM_K})^T \quad , \quad x_n = (x_{11},\ldots,x_{km_k})^T$$

be respectively the $N \times q$ matrix of finite population values and $n \times q$ observed data matrix discussed in Section 1, where T denotes transpose.

For simplicity we make the following assumption in Sections 2 - 4.

Assumption 1: There is no auxiliary information to distinguish the clusters, in particular the cluster sizes are equal: $M_i = M$, $i = 1,\ldots,K$.

We consider the case of unequal M_i in Section 5. We now define the distributions π_0 and π_1, of x_n.

Definition 2.1: The true distribution of x_n, denoted by $\pi_1(x_n)$, obeys the following conditions:

(i) conditional on (random) distribution functions, $F_1,\ldots,F_k$, the x_{ij} are mutually independent and

$$x_{ij} \mid F_1,\ldots,F_k \sim F_i \qquad i = 1,\ldots,k;\; j = 1,\ldots,m_i \quad ,$$

(ii) $F_1,\ldots,F_k$ are IID.

Remark: In (ii) the F_i are functions on R^q and so the distribution of each F_i is infinite dimensional as in the theory of stochastic processes. More precistely we might follow Ferguson (1974) and let Φ be a set of distribution functions on R^q, Θ be a sigma-algebra of subsets of Φ and Π be a probability measure on (Φ,Θ). Then, equivalently to (ii), we assume $(F_1,\ldots,F_k)$ is an outcome of the product space $(\Phi,\Theta,\Pi)^k$.

Design-based Interpretation of π_1.

This distribution can be viewed as the randomization distribution of x_n induced by simple random sampling with replacement at both stages. Let G_α be the 'empirical' distribution function of x in the α^{th} cluster, i.e. G_α assigns probability mass M^{-1} to each point $x_{\alpha 1},\ldots,x_{\alpha M}$. Let $\Phi = \{G_1,\ldots,G_K\}$ and let Π assign probability K^{-1} to each outcome G_α. Hence each F_i is equal to a randomly chosen G_α.

Model-based interpretation of π_1.

Suppose x is a realization of the $N{\times}p$ random matrix, X, with prior distribution $\pi_1(x)$ obtained by extending Definition 2.1 by substituting K for k and M for m_i. Suppose that the sample design, $p(S)$, is non-informative in the sense that S and X are independent. Then $\pi_1(x_n)$ is the appropriate distribution of x_n for model-based inference conditional on $S = s$ (Sugden and Smith, 1984). This is inference procedure (A) referred to in Section 1.

The distribution $\pi_1(x)$ seems both a natural and a general non-parametric model for expressing the symmetry between clusters and between units within clusters. A simple example is the one-way random effects model (e.g. Scott and Smith, 1969). Here Φ is a

location family, $\Phi = \{G(x-\varphi); \varphi \in R^q\}$ where G is a given distribution function and Π defines a prior distribution for φ. For example, Φ may correspond to the normal family $\{N_q(\varphi,\Omega_W); \varphi \in R^q\}$ and Π may correspond to $\varphi \sim N_q(\mu,\Omega_B)$. Other examples with scale parameters and higher-order cumulants varying between clusters are given by Leonard (1975) and Skinner (1981) respectively.

The distribution $\pi_1(x)$ is also a special case of the two-stage exchangability/random permutation model of Bellhouse et al (1977). Their model is more general because, in particular, it allows for negative intra-cluster correlation as does the similar model of Royall (1976). However, it is less interpretable and, for example, would not permit Theorem 3.6., one of our main results. Furthermore, if we add the assumption that x is part of a doubly infinite sequence $\{x_{ij}; i = 1,2,\ldots, j = 1,2,\ldots\}$ such that (i) and (ii) hold for any K and M then we would conjecture that this two-stage exchangeability model could be represented by π_1. (Aldous, 1981, proves a stronger result for a crossed rather than a nested doubly infinite array.)

Definition 2.2: The baseline distribution of x_n, denoted by $\pi_0(x_n)$, obeys the following condition:

(i) $x_{11},\ldots,x_{km_k}$ are IID.

Design-based interpretation of π_0.

This is the randomization distribution induced in x_n by simple random sampling with replacement from the whole finite population.

Model-based interpretation of π_0.

This is the 'textbook' IID assumption referred to in Section 1.

We assume the existence of the first two moments of x_n under both π_0 and π_1 and write

$$E_{\pi_1}(x_{ij}) = \mu \tag{2.1}$$

$$\begin{aligned} \mathrm{cov}_{\pi_1}(x_{ij}, x_{i'j'}) &= \Omega \qquad i = i',\ j = j' \\ &= \Omega_B \qquad i = i',\ j \neq j' \\ &= 0 \qquad i \neq i' \ . \end{aligned}$$

Note that μ and Ω are the finite population mean and covariance matrix under the design-based interpretation.

Finally, given the nature of Definition 1.1 we need to define an asymptotic framework. We follow Fuller (1975, Appendix A) in considering a sequence of finite populations and designs $\{U_k, p_k;\ k = 1,2,\ldots\}$ such that U_k contains K_k clusters, $K_k > K_{k-1}$, and p_k selects k clusters of sizes $m_1,\ldots,m_k$ from U_k. We assume $\{m_1, m_2,\ldots\}$ is a fixed infinite sequence with $1 \leq m_i \leq M$. The limits $k \to \infty$ and $n = \sum^{k} m_i \to \infty$ are then equivalent but we use the latter notation to be consistent with Section 1. We assume that Definitions 2.1 and 2.2 can be extended for $k = 1,2,\ldots$ and that the common distribution of F_i in Definition 2.1 and the common distribution of x_{ij} in Definition 2.2 does not depend on k. From the design-based interpretation this implies further restrictions if π_0 and π_1 are not to depend on k via K_k. One approach (c.f. Brewer, 1979) is to assume that K_k is an integer multiple of k, say $K_k = Lk$. Then suppose that $x_{ij} = x_{i_0 j}$, $i = L+1,\ldots,K$; $j = 1,\ldots,M$ where $i_0 = (i-1) \bmod L + 1$, that is x consists of k 'reproductions' of the $LM \times p$ matrix $(x_{11},\ldots,x_{LM})$. Then $\Phi = \{G_1,\ldots,G_L\}$ and Π, which assigns probability L^{-1} to each G_α, $\alpha = 1,\ldots,L$, no longer depends on K_k. Alternatively one might introduce super-population assumptions as in Fuller (1975, Appendix A).

3. RESULTS

We consider the class of estimators of form

$$t_n = g(\bar{x}_n) \tag{3.1}$$

where $\bar{x}_n = n^{-1} \sum_s x_{ij}$ and g is a given function $g : R^q \to R$. This class includes, in particular, functions of second moments such as correlation coefficients and linear regression coefficients by defining x to include squares and products of the 'raw' survey variables (see, for example, Section 4 and Krewski and Rao, 1981). For simplicity we assume g is scalar-valued but results extend straightforwardly to vector-valued g. We assume the target parameter is

$$\theta_1 = g(\mu) \qquad (3.2)$$

where μ is defined in (2.1). For example, if t_n is the sample correlation coefficient then θ_1 is the finite population correlation coefficient under the design-based interpretation or the 'super-population' correlation coefficient under the model-based interpretation of π_1.

The main aim of this section is to apply Definition 1.1 to t_n in (3.1) for the model π_1 of Definition 2.1. This will be done in Theorem 3.5 but first we need to establish the conditions of Definition 1.1 for t_n and π_1. We make use of the following

Condition $C1(\pi_1)$: For some $\varepsilon > 0$, $E_{\pi_1} |(x_{ij} - \mu)_\ell|^{2+\varepsilon}$ exists for $\ell = 1,\ldots,q$, where $(\cdot)_\ell$ denotes the ℓ^{th} element of a vector.

Condition $C2(\pi_1)$: The function g admits continuous partial derivatives at μ at least one of which does not vanish at μ.

Conditions $C1(\pi_0)$ and $C2(\pi_0)$ are defined analogously with π_0 and $E_{\pi_0}(x_{ij})$ replacing π_1 and μ respectively.

The corollary of the following lemma establishes one conditon of Definition 1.1 and gives the numerator of (1.1).

Lemma 3.1

If $C1(\pi_1)$ holds then under π_1 as $n \to \infty$

$$n^{1/2}(\bar{x}_n - \mu) \xrightarrow{L} N_q[0, (I_q + (m^*-1)\Gamma)\Omega] \qquad (3.3)$$

where $\Gamma = \Omega_B \Omega^{-1}$, $m^* = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{k} m_i^2$.

Proof: Let $z_i = kn^{-1} \sum_{j=1}^{m_i} (x_{ij} - \mu)$. Then $\{z_1, z_2 \ldots\}$ is a sequence of independent random

vectors with

$$E_{\pi_1}(z_i) = 0 \quad , \; E_{\pi_1}(z_i z_i^T) = k^2 n^{-2} m_i (I_q + (m_i - 1)\Gamma)\Omega \quad ,$$

$$E_{\pi_1}|(z_i)_\ell|^{2+\varepsilon} = (kn^{-1})^{2+\varepsilon} E_{\pi_1} \Big| \sum_{j=1}^{m_i} (x_{ij} - \mu)_\ell \Big|^{2+\varepsilon}$$

$$< m_i^{2+\varepsilon} E_{\pi_1} |(x_{ij} - \mu)_\ell|^{2+\varepsilon} \quad \text{by Minkowski's inequality and since } n > k$$

$$= O(1) \quad \text{from } C1(\pi_1) \text{ and since } m_i < M \quad .$$

Hence by using a central limit theorem such as Lemma 3.1 of Krewski and Rao (1981)

$$k^{1/2}(k^{-1} \sum_{i=1}^{k} z_i) \xrightarrow{L} N_q(0, \sigma_z^2)$$

$$\text{where} \quad \sigma_z^2 = \lim_{n\to\infty} k^{-1} \sum_{i=1}^{k} E(z_i z_i^T)$$

$$= \lim_{n\to\infty} kn^{-1}(I_q + (n^{-1}\Sigma m_i^2 - 1)\Gamma)\Omega \quad .$$

The result follows since $\bar{x} - \mu = k^{-1}\Sigma z_i$.

From standard asymptotic theory we obtain the following:

Corollary 3.2

If $C1(\pi_1)$ and $C2(\pi_1)$ hold then under π_1 as $n \to \infty$

$$n^{1/2}(t_n - \theta_1) \xrightarrow{L} N[0, \; (1 + (m^* - 1)\rho_g)\sigma_g^2] \tag{3.4}$$

$$\text{where} \quad \sigma_g^2 = \nabla_g(\mu)^T \Omega \nabla_g(\mu) \quad , \quad \nabla_g(\mu) = \partial g(\mu)/\partial\mu$$

$$\rho_g = \nabla_g(\mu)^T \Omega_B \nabla_g(\mu)/\sigma_g^2 \quad . \tag{3.5}$$

Remarks

1. Under the model-based approach Corollary 3.2 would also hold if $\theta_1 = g(\bar{x}_N)$ where $\bar{x}_N = N^{-1} \sum_U x_{ij}$ provided $n/N \to 0$ as $n \to \infty$. Fuller (1975, Appendix A) gives a result

corresponding to Lemma 3.1 for $\bar{x}_n - \bar{x}_N$ where $n/N \to f \neq 0$.

2. m^* exists because the m_i are bounded, $1 \leq m_i \leq M$.

3. The quantity ρ_g is the intra-cluster correlation of

$$w_{ij} = \nabla_g(\mu)^T x_{ij} \tag{3.6}$$

since
$$\rho_g = \mathrm{corr}_{\pi_1}(w_{ij}, w_{ij'}),\ j \neq j' \quad . \tag{3.7}$$

We may write alternatively

$$\rho_g = \mathrm{var}_{\pi_1}(\nabla_g(\mu)^T \mu_i)/\mathrm{var}_{\pi_1}(w_{ij}) \tag{3.8}$$

where
$$\mu_i = E_{\pi_1}(x_{ij}|F_i) = \int x dF_i(x) \quad . \tag{3.9}$$

The delta-method or Taylor-series linearization estimator of the variance of t_n under the assumption that π_0 is true is $n^{-1}v_{g,0,n}$ where

$$v_{g,0,n} = \nabla_g(\bar{x}_n)^T v_{0,n} \nabla_g(\bar{x}_n) \tag{3.10}$$

$$v_{0,n} = (n-1)^{-1} \sum_s (x_{ij} - \bar{x}_n)(x_{ij} - \bar{x}_n)^T \quad .$$

The corollary of the following lemma establishes another condition of Definition 1.1 and gives the denominator of (1.1).

<u>Lemma 3.3</u>

If $C1(\pi_1)$ holds then as $n \to \infty$

$$v_{0,n} \xrightarrow{\pi_1} \Omega$$

where $\xrightarrow{\pi_1}$ denotes convergence in probability under π_1.

<u>Proof</u>: We may write

$$v_{0,n} = (n-1)^{-1} k [k^{-1} \sum_{i=1}^{k} u_i] + (n-1)^{-1} n\Omega - n(n-1)^{-1}(\bar{x}-\mu)(\bar{x}-\mu)^T$$

where
$$u_i = \sum_{j=1}^{m_i} [(x_{ij}-\mu)(x_{ij}-\mu)^T - \Omega] \quad .$$

Now $\{u_1, u_2, \ldots\}$ is a sequence of zero-mean independent random matrices. If the

m_i are equal the u_i are IID and by Khinchine's version of the Weak Law of Large Numbers $k^{-1}\Sigma u_i \xrightarrow{\pi_1} 0$ even without $C1(\pi_1)$. If the m_i are unequal then the application of Minkowski's inequality as in Lemma 3.1 and the use of $C1(\pi_1)$ together with the fact that $m_i < M$ implies that the $(1+\frac{1}{2}\varepsilon)^{th}$ moment of the absolute value of each element of u_i is bounded uniformly in i and so (e.g. Krewski and Rao, 1981, Lemma 3.2) $k^{-1}\Sigma u_i \xrightarrow{\pi_1} 0$. The result then follows by noting that $(n-1)^{-1}k$ is bounded and that $(\bar{x}_n-\mu) \xrightarrow{\pi_1} 0$ from Lemma 3.1.

From the assumed continuity of ∇_g and the fact that $\bar{x}_n \xrightarrow{\pi_1} \mu$ we obtain:

<u>Corollary 3.4</u>

If $C1(\pi_1)$ and $C2(\pi_1)$ hold then as $n \to \infty$

$$v_{g,0,n} \xrightarrow{\pi_1} \sigma_g^2 \quad . \tag{3.11}$$

We are now in a position to derive our main result.

<u>Theorem 3.5</u>

If $C1(\pi_0)$, $C1(\pi_1)$, $C2(\pi_0)$, $C2(\pi_1)$ hold then

$$\text{deff}[t_n;\ \pi_1,\ v_{g,0,n}] = 1 + (m^*-1)\rho_g \quad . \tag{3.12}$$

<u>Proof</u>: The condition of Definition 1.1 hold from Corollaries 3.2 and 3.4 and by noting that π_0 is a special case of a model of form π_1 with $F_1 = \ldots = F_k$. The expression in (3.12) is obtained from (3.4) and (3.11).

<u>Remarks</u>

1. If $m_1 = \ldots = m_k = m$ then $m^* = m$ and the expression in Theorem 3.5 has the familiar form of the design effect of a mean (Kish, 1965). The IID-based estimator $v_{g,0,n}$ under-estimates the variance of t_n by an amount which depends on the subsample size m and the intra-cluster correlation ρ.

2. If the m_i are unequal note that

$$m^* = \lim_{n\to\infty}[\bar{m} + \Sigma(m_i-\bar{m})^2/n] \geq \lim_{n\to\infty} \bar{m} \quad \text{where} \quad \bar{m} = n/k \quad .$$

Hence expression (3.12) tends to be greater than the commonly used expression $1 + (\bar{m}-1)\rho_g$ (Kish, 1965). Our expression for m^* is the limit of expressions appearing in Campbell (1977) and Rao and Scott (1981).

3. Referring to Remark 2 under Definition 1.1 the asymptotic relative bias is zero under the design-based interpretation since $\theta_0 = \theta_1 = g(\bar{x}_N)$ and should be negligible under the model-based interpretation because $t_n - g(\bar{x}_N) \overset{P}{\to} 0$ under either $\pi = \pi_0$ or $\pi = \pi_1$.

For the remainder of this section we examine the quantity ρ_g in (3.12). We may view ρ_g either as an intra-cluster correlation of w_{ij} as in (3.7) or as a measure of homogeneity of the $\nabla_g(\mu)^T\mu_i$ as in (3.8). For example, if $q = 1$ and g is the identity function then t_n is the sample mean $\bar{x}_n$ and ρ_g is the usual intra-cluster correlation of the x_{ij} which is a measure of homogeneity of the means μ_i in the different clusters. In general, however, neither (3.7) or (3.8) are very easy to interpret because of their dependence on the rather artificial quantities w_{ij} and $\nabla g(\mu)^T\mu_i$. In order to obtain a more interpretable expression for ρ_g we impose a further condition on the distribution π_1. This condition is strictly only applicable under the model-based approach.

Referring to Definition 2.1 let $F = E_{\pi_1}(F_i)$ be the marginal distribution of x_{ij}. We suppose each F_i is a mixture.

<u>Condition C3:</u> $$F_i = (1-\delta)F + \delta D_i \qquad i = 1,\dots,k$$

where $0 < \delta < 1$ and $D_1,\dots,D_k$ are IID distribution functions with $E(D_i) = F$.

One extreme $\delta = 0$ then corresponds to π_0 whilst the other extreme $\delta = 1$ imposes no further structure on π_1. We shall suppose that δ is small which we suggest is a natural non-parametric way of asserting that there is low intra-cluster correlation. This assumption may not be unreasonable in, for example, large-scale sample surveys where the clusters are geographical areas. In such surveys the intra-cluster correlation of variables is usually low (say < 0.1) by design, even though the design effect may be non-negligible because of the value of m^*.

We need further regularity conditions.

<u>Condition C4:</u> The matrix H_g of second partial derivatives of g exists in a neighborhood of μ and

$$\mathrm{var}_{\pi_1}[(\mu(D_i)-\mu)^T H_g[\mu+\varepsilon(\mu(D_i)-\mu)](\mu(D_i)-\mu)]$$

is bounded as $\varepsilon \downarrow 0$ where we use the functional notation $\mu(D_i) = \int x dD_i(x)$ so that $\mu_i = \mu(F_i)$, $\mu = \mu(F)$.

The following theorem gives an alternative approximate expression for ρ_g when the intra-cluster correlation is low.

<u>Theorem 3.6.</u>

If C3 and C4 hold then as $\delta \downarrow 0$

$$\rho_g = \mathrm{var}_{\pi_1}[g(\mu_i)]/\mathrm{var}_{\pi_1}(w_{ij}) + O(\delta^3) = O(\delta^2)$$

where μ_i and w_{ij} are defined in (3.9) and (3.6).

<u>Proof:</u> Consider the Taylor Series expansion

$$g(\mu_i) = g(\mu) + \nabla_g^T(\mu)(\mu_i-\mu) + \frac{1}{2}(\mu_i-\mu)^T H_g(\mu^*)(\mu_i-\mu) \tag{3.13}$$

where $\mu^* = (1-\phi)\mu + \phi\mu_i$ and ϕ is a scalar, $0 < \phi < 1$. Now

$$\begin{aligned} \mu_i-\mu &= \mu(F_i) - \mu(F) = \mu(F_i-F) \\ &= \mu[\delta(D_i-F)] \quad \text{from C3} \\ &= \delta[\mu(D_i)-\mu] \quad . \end{aligned} \tag{3.14}$$

The result follows by substituting (3.14) into (3.13) and using (3.8) and C4 with $\varepsilon = \phi\delta$.

The quantity $g(\mu_i)$ is the cluster 'version' of $t_n = g(\bar{x}_n)$ and $\theta_1 = g(\mu)$. For example, if t_n is the sample correlation coefficient and θ_1 is the population correlation coefficient then $g(\mu_i)$ is the correlation coefficient in the i^{th} cluster. A specific example is given in Section 4 where t_n is the sample variance and $g(\mu_i)$ is the variance in the i^{th} cluster. The quantity $\mathrm{var}(w_{ij})$ does not depend on the clustering in the population (in terms of C3 it depends on F but not on δ or D_i) and may be viewed as a standardizing quantity. Hence Theorem 3.6 permits ρ_g to be interpreted as a measure of homogeneity of the quantities $g(\mu_i)$, providing the overall level of intra-cluster correlation is 'low'. Combining with Theorem 3.5 suggests, for example, that the design effect of a sample correlation coefficient is mainly determined by the difference between the correlation coefficients within clusters.

4. AN EXAMPLE: DESIGN EFFECT OF A SAMPLE VARIANCE

The low-δ approximation in Theorem 3.6 is examined here explicitly for the case where

$$t_n = n^{-1} \sum_s (y_{ij}-\bar{y}_n)^2 \quad , \quad \bar{y}_n = n^{-1} \sum_s y_{ij} \ .$$

We may write t_n in the form of (3.1) by letting $q = 2$, $x_{ij} = (y_{ij}, y_{ij}^2)^T$, $g[(x_1,x_2)^T] = x_2 - x_1^2$. Hence $\bar{x}_n = (\bar{y}_n, n^{-1} \sum y_{ij}^2)$, $g(\bar{x}_n) = n^{-1} \sum y_{ij}^2 - \bar{y}_n^2 = t_n$. Following (2.1) and (3.9) define the within-cluster and overall moments by

$$\mu_i = (\mu_{yi}, \mu_{yi}^2 + \sigma_{yi}^2)^T \quad , \mu = (\mu_y, \mu_y^2 + \sigma_y^2)^T \ .$$

Then $g(\mu_i) = (\mu_{yi}^2 + \sigma_{yi}^2) - \mu_{yi}^2 = \sigma_{yi}^2$ is the within-cluster variance corresponding to the sample variance t_n and the population variance $g(\mu) = \sigma_y^2$. Also $\nabla_g(\mu) = (-2\mu_y, 1)$ obeys $C2(\pi_1)$ and from (3.6), up to an additive constant

$$w_{ij} = (y_{ij} - \mu_y)^2 \ .$$

The IID-based variance estimate of t_n given by (3.10) is

$v_{g,0,n} = (n-1)^{-1} \sum_s (w_{ij}^* - \bar{w}^*)^2$ where $w_{ij}^* = (y_{ij} - \bar{y}_n)^2$, $\bar{w}^* = t_n$.

The low-δ approximation to ρ_g given by Theorem 3.6 is

$$\rho_g^+ = \mathrm{var}_{\pi_1}[g(\mu_i)]/\mathrm{var}_{\pi_1}(w_{ij}) = \mathrm{var}_{\pi_1}(\sigma_{yi}^2)/\mathrm{var}_{\pi_1}(y_{ij}-\mu_y)^2$$

which may be compared with the expression from (3.8)

$$\rho_g = \mathrm{var}_{\pi_1}[\sigma_{yi}^2 + (\mu_{yi}-\mu_y)^2]/\mathrm{var}_{\pi_1}(y_{ij}-\mu y)^2 \ .$$

Hence we may write

$$\rho_g^+ < \rho_g < \rho_g^+ + 2(\eta\rho_g^+)^{1/2} + \eta$$

where $\eta = \mathrm{var}_{\pi_1}[(\mu_{yi}-\mu_y)^2]/\mathrm{var}_{\pi_1}[(y_{ij}-\mu_y)^2]$.

Define the between-cluster and total coefficients of kurtosis by

$\gamma_B = E_{\pi_1}(\mu_{yi}-\mu_y)^4/[\mathrm{var}_{\pi_1}(\mu_{yi})]^2 - 3$, $\gamma = E_{\pi_1}(y_{ij}-\mu_y)^4/[\mathrm{var}_{\pi_1}(y_{ij})]^2 - 3$.

Let $\mathrm{var}(\mu_{yi})/\mathrm{var}(y_{ij})$ be the conventional intra-cluster correlation of y_{ij}. Then

$$\eta = \frac{(2+\gamma_B)}{(2+\gamma)} \rho_y^2 \ . \tag{4.1}$$

Hence if ρ_g is small then η will be very small unless γ is very small (for example $\gamma = -1.2$ for the very platykurtic uniform distribution) or γ_B is very large (for example $\gamma_B = 6$ for the very leptokurtic exponential distribution). Thus if ρ_g is small and there is reasonable dispersion amongst the σ_i^2 then $\rho_g \doteq \rho_g^+$ should be a fair

approximation. In terms of δ, both $\mu_{yi} - \mu_y$ and $\sigma_{yi}^2 - E(\sigma_{yi}^2)$ are of $O_{\pi_1}(\delta)$, ρ_g^+ and ρ_y are of $O(\delta^2)$ whilst η is of $O(\delta^4)$.

5. UNEQUAL CLUSTER SIZES

The results in Sections 3 and 4 were based on the assumption of equal cluster sizes. If the M_i are unequal and Definition 2.1 (which does not involve the M_i) still applies then these results will still hold (provided $\{m_1, m_2, \ldots\}$ is a fixed bounded sequence).

From the design-based viewpoint, Definition 2.1 still holds under simple random sampling with replacement at both stages where the m_i are fixed and do not depend on the M_i.

From the model-based viewpoint, Definition 2.1 remains appropriate if the within-cluster distributions F_i do not depend on the M_i. It does not matter here if the design $p(S)$ is dependent on the M_i as for example in probability proportional to size sampling.

In general F_i may depend on M_i and the results of Section 3 will not hold. For example, $\bar{x}_n$ may no longer even be a consistent estimator of μ. A general discussion of inference under models for populations with unequal size clusters is given by Sundberg (1983). We suggest, however, that within strata, and in particular within size strata, our results should hold at least approximately. In fact, plots of $\bar{x}_i = m_i^{-1} \sum_{j=1}^{m_i} x_{ij}$ and $m_i^{-1} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i)^2$ against M_i for various variables x and data sets in Skinner (1982) suggested little relation between F_i and M_i.

6. DISCUSSION

Under given conditions, in particular when the number of sampled clusters is large, the design effect of two-stage sampling was shown in Theorem 3.5 to take the familiar form, $1 + (m-1)\rho$, for a broad class of statistics. This result has an interpretation both from the design-based viewpoint in terms of with replacement sampling and also from a model-

based viewpoint in terms of a fairly general non-parametric model for a clustered population.

For linear statistics, such as the sample mean, ρ may be interpreted as a measure of homogeneity of corresponding within cluster quantities, such as cluster means. For non-linear statistics, such as the sample correlation coefficient, provided the overall level of intracluster correlation is not high, it was shown in Theorem 3.6 that ρ may also be interpreted as a measure of homogeneity of corresponding within cluster quantities, such as cluster correlation coefficients.

These results have rather negative implications for the existence of relations between design effects of multivariate and of univariate statistics as discussed at the end of Section 1. In general we conclude no necessary theoretical relation need hold. For example, the design effect of a correlation coefficient, being determined mainly by the heterogeneity of cluster correlations, has in general no necessary relation with the design effects of the means of the two variables, which are determined by the heterogeneity of the cluster means. Our conclusion agrees with that of Rao and Scott (1981) on the design effect involved in testing independence in a bivariate contingency table. They state that 'ideally we would like an approximation ... based on the marginal design effects' (that is the univariate design effects) but 'such an approximation does not seem possible in theory'.

Theoretical relations can be derived under restricted assumptions but such results can be misleading. For example, a regression model of y on z with errors correlated within clusters but regression slopes β constant across clusters is considered by Campbell (1977) and Scott and Holt (1982). They obtain the $1 + (m-1)\rho$ result for the least-squares estimator of the slope and show that $\rho = \rho_z \rho_e$ where ρ_z and ρ_e are the intra-cluster correlations of z and of the residual $e = y-\beta z$ respectively. Now if both ρ_z and ρ_e are small then ρ is very small which the authors take to correpond to Kish and Frankel's (1974) empirical observation that 'design effects for complex statistics tend to be less than those for means of the same variable'. However, this approach effectively assumes away the dominating $O(\delta^2)$ term in Theorem 3.6 determined by the dispersion-

between cluster regression coefficients and just obtains the $O(\delta^4)$ term analogous to η in (4.1). Hence we suggest the above formula could drastically underestimate the true design effect. Other examples of the application of Theorems 3.5 and 3.6 for specific statistics and under restricted assumptions are given in Skinner (1982).

Rao and Scott (1981), following on from their statement above, suggest that 'it may be possible to find empirically-based approximations that work well in practice'. In another context, for example, Bebbington and Smith (1977) suggest an empirical relation between the design effect of a correlation coefficient and the minimum of the design effects of the corresponding means. The derivation of such empirical 'laws', whilst potentially useful, is no easy project without guidance from theory, given the infinite range of possible statistics, designs and population structures.

REFERENCES

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. J. Multivariate Anal. 11, 581-598.

BEBBINGTON, A. C. and SMITH, T. M. F. (1977). The effect of survey design on multivariate analysis. In The Analysis of Survey Data Vol. 2 (C. A. O'Muircheartaigh and C. Payne, eds). New York: Wiley.

BELLHOUSE, D. R., THOMPSON, M. E. and GODAMBE, V. P. (1977). Two-stage sampling with exchangeable prior distributions. Biometrika, 64, 97-103.

BREWER, K. R. W. (1979). A class of robust sampling designs for large-scale surveys, J. Amer. Statist. Ass. 74, 911-915.

CAMPBELL, C. (1977). Properties of ordinary and weighted least-squares estimators of regression coefficients for two-stage samples. ASA: Proc. Soc. Stat. Sect. 800-805.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. Ann. Statist. 2, 615-629.

FULLER, W. A. (1975). Regression analysis for sample surveys. Sankhyā C 37, 117-132.

KISH, L. (1965). Survey Sampling. New York: Wiley.

KISH, L. and FRANKEL, M. R. (1974). Inference from complex samples. J. R. Statist. Soc. B 36, 1-37.

KREWSKI, D. and RAO, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist. 9, 1010-1019.

LEONARD, T. (1975). A Bayesian approach to the linear model with unequal variances. Technometrics 17, 95-102.

NATHAN, G. and HOLT, D. (1980). The effect of survey design on regression analysis. J. R. Statist. Soc. B 42, 377-386.

PFEFFERMAN, D. and NATHAN, G. (1981). Regression analysis of data from a cluster sample. J. Amer. Statist. Ass. 76, 681-689.

RAO, J. N. K. and SCOTT, A. J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. J. Amer. Statist. Asso, 76, 221-230.

ROYALL, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. J. Amer. Statist. Ass., 71, 657-664.

ROYALL, R. M. and CUMBERLAND, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. J. Amer. Statist. Ass., 76, 66-88.

SARNDAL, C-E. (1978). Design-based and model-based inference in survey sampling. Scand. J. Statist., 5, 27-52.

SCOTT, A. J. and HOLT, D. (1982). The effect of two-stage sampling on ordinary least-squares methods. J. Amer. Statist. Ass., 77, 848-854.

SCOTT, A. J. and SMITH, T. M. F. (1969). Estimation in multi-stage surveys. J. Amer. Statist. Ass., 68, 880-889.

SKINNER, C. J. (1981). Estimation of the variance of a finite population or cluster samples. Sankhyā, B, 392-398.

SKINNER, C. J. (1982). Multivariate analysis of sample survey data. Unpublished Ph.D. Thesis. University of Southhampton.

SUGDEN, R. A. and SMITH, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. To appear in Biometrika.

SUNDBERG, R. (1983). The predictive approach and randomized population type models for finite population inference from two-stage samples. Scand. J. Statist., 10, 223-238.

CJS/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2841	2. GOVT ACCESSION NO. AD-A160960	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Design Effects of Two-Stage Sampling		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) C. J. Skinner		8. CONTRACT OR GRANT NUMBER(s) DMS-8210950, Mod. 1 DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53705		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS See Item 18 below		12. REPORT DATE July 1985
		13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Design Effect; Model Misspecification; Two-stage Sampling; Finite Population; Sample Survey

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The effect of a two-stage sampling design on statistical inference is discussed. A definition of a design effect is given. The structure of design effects for a class of statistics is investigated. Results have both a design-based and a model-based interpretation. The relation between design effects for multivariate statistics and design effects for univariate statistics is considered.

END

FILMED

12-85

DTIC